Data 6 Summer 2026

Abby Brooks-Ramirez

Quiz 02

Your name: _____

Your student ID: _____

Your Berkeley email: _____

Your room location: _____

Student ID of the person to your left: _____

Student ID of the person to your right: _____

You have 50 minutes. There are 3 questions of varying credit. (35 points total)

Question:	HC	1	2	3	Total
Points:	1	13	21	0	35

For questions with **circular bubbles**, you may select only one choice.

- ○ Unselected option (Completely unfilled)
- ◎ Don't do this (it will be graded as incorrect)
- ● Only one selected option (completely filled)

For questions with **square checkboxes**, you may select one or more choices.

- ■ You can select
- ■ multiple squares
- ☑ (Don't do this)

Anything written outside the answer boxes or ~~crossed out~~ will not be graded. If you write multiple answers, your answer is ambiguous, or the bubble/checkbox is not entirely filled in, we will grade the worst interpretation. For coding questions with blanks, you may write at most one statement per blank, unless otherwise stated.

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will follow the rules of this exam.

Honor Code (HC): I have read and agree to the honor code above.

(1 point) Sign your name: _____

Q1 *Li-Bear-ies!* **(13 points)**

Suppose we want to settle the debate–which UC Berkeley library is the most popular? We have created **libraries**, a table that measures how many people are occupying each library based on the time of day and number of weeks into the semester (Table 1). One row represents one measurement of a library's occupants at a certain time, week number, and library. Assume there is exactly 1 occupancy measurement for each combination of week, library, and time.

week	library	time of day	occupants
2	Doe Library	Evening	528
1	Main Stacks	Afternoon	307
3	Moffitt Library	Morning	438
1	Doe Library	Morning	295
3	Main Stacks	Evening	435
2	Moffitt Library	Afternoon	531

Table 1: Six random rows of the **libraries** table.

Variable descriptions:
- **week**: The week number the library occupancy was measured in (integer).
- **library**: The library whose occupancy was measured (string).
- **time of day**: The time of day the occupancy was measured (string).
- **occupants**: The number of library occupants at that specific library, time, and week (integer).

Q1.1 (1 point) Which of the below would be best for visualizing how the average occupancy across all libraries changes over the course of the semester?

◯ Bar chart ◯ Scatter plot ◯ Box plot

◯ Histogram ◯ Line plot ◯ None of the above

Q1.2 (1 point) Which of the below would be best for comparing average occupancy across different times of day?

◯ Bar chart ◯ Scatter plot ◯ Box plot

◯ Histogram ◯ Line plot ◯ None of the above

Q1.3 (2 points) Which of the below would allow us to compare the distribution of library occupancy between different libraries? Select all that apply.

☐ Bar chart ☐ Scatter plot ☐ Box plot

☐ Histogram ☐ Line plot ☐ Pie Chart

Consider the **means** table below, which contains the average occupancy for each unique library in the **libraries** table.

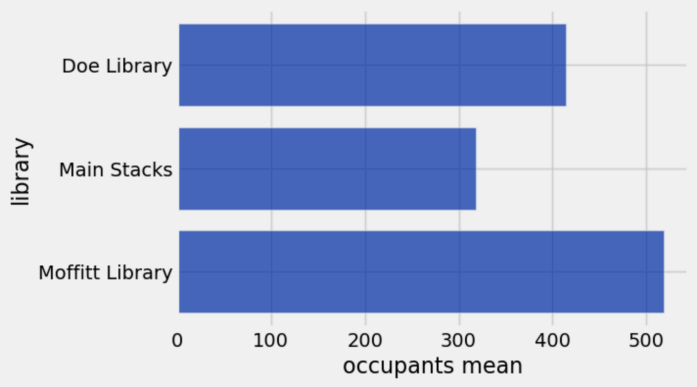
library	occupants mean
Doe Library	414.6
Main Stacks	317.9
Moffitt Library	518.8

Table 2: **means** (Mean occupancy by library)

Q1.4 (2 points) Which of the below lines of code will assign **means** to the table above? Select all that apply.

- ☐ **means = libraries.group("library", np.mean)**
- ☐ **means = libraries.select("library", "occupants").group("library", np.mean)**
- ☐ **means = libraries.group("library", np.mean).select("library", "occupants mean")**
- ☐ **means = libraries.pivot("library", "occupants", np.mean).select("library", "occupants mean")**

Q1.5 (2 points) Assume **means** is correctly assigned above. Which line of code produces the following visualization? Select all that apply.



- ☐ **means.barh("library", "occupants mean")**
- ☐ **means.barh("library")**
- ☐ **means.hist("library")**
- ☐ **means.hist("library", "occupants mean")**

Your friend Oski looks at the visualization from question 1.5 and claims that Moffitt Library is the most popular library at UC Berkeley. Since it has the highest average occupancy, Oski argues it must have the highest occupancy at any time you check it.

You want to investigate this claim by looking at average occupancy broken down by time of day. Fill in the code below to create the table **occupancy_by_time**, included below.

time of day	Doe Library	Main Stacks	Moffitt Library
Afternoon	487.625	500	421.9
Evening	300	450	374.4
Morning	250	340.625	336.9

Table 3: The **occupancy_by_time** table.

```
occupancy_by_time = libraries.____(a)____(_________(b)_________)
```

Q1.6 (1 point) Fill in blank (a)

Q1.7 (2 points) Fill in blank (b)

Q1.8 (2 points) Based on the table **occupancy_by_time** and the visualization in Q1.5, how would you respond to Oski's claim? You must cite Simpson's paradox in our response. Please limit your response to 2-3 sentences.

Q2 *Running the Numbers* **(21 points)**

The **marathon** dataset contains the finishing times of the runners who competed in the 2019 Boston Marathon. Each row represents a runner who completed the race, and over 26,000 people completed the race in total!

rank	age	gender	country	time
24615	44	M	Canada	18483
12587	35	F	Turkey	13374
18478	41	F	United States	14978
21048	26	M	United States	15988
7562	51	M	India	12250

Table 4: The first 5 rows of the **marathon** table (26,641 rows in total).

Variable descriptions:

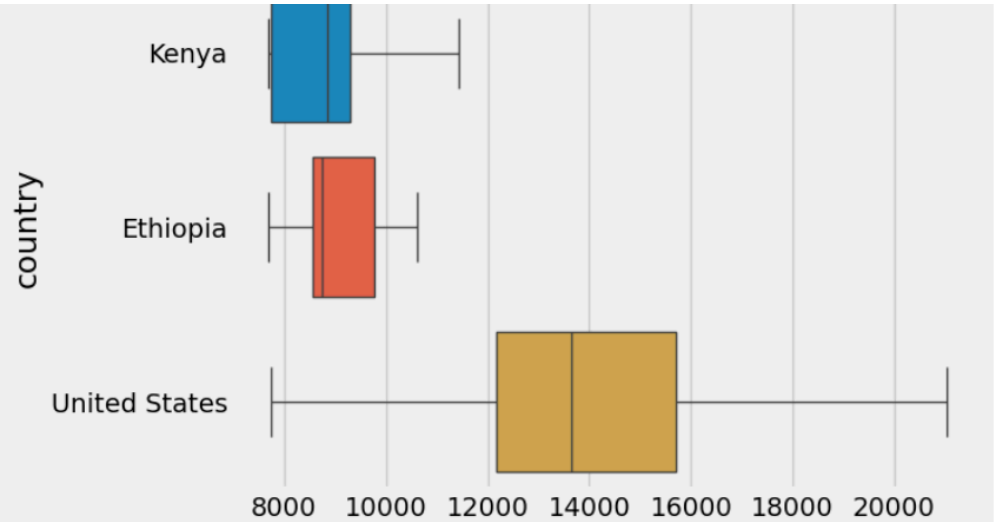
- **rank**: Overall finishing position of the runner among all finishers (integer).
- **age**: Age of the runner in years at the time of the race (integer).
- **gender**: Gender of the runner (string).
- **country**: Name of the country the runner represents (string).
- **time**: Finishing time of the runner (the total amount of time taken to complete the marathon), measured in seconds (integer).

Q2.1 (2 points) In 2019, the countries with the most runners in the top 10 were Kenya, Ethiopia, and the United States. Fill in the code below to assign **filtered_countries** to a table containing only runners representing the countries in the array **countries**.

```
countries = make_array('Kenya', 'Ethiopia', 'United States')

filtered_countries = marathon._____
                               Q2.2
```

Using our new table **filtered_countries** we can create the following box plot, with the x-axis representing **time**. We will use this box plot to answer the next two questions:



Q2.3 (2 points) Using only the box plots in the image above, which of the following statements are **true**?

- ☐ The middle 50% of Ethiopian runners had shorter finishing times than the middle 50% of United States runners
- ☐ There are fewer runners from Kenya than runners from the United States
- ☐ The interquartile range of United States runners is larger than the interquartile range of Kenyan runners
- ☐ On average, Ethiopian runners had shorter finishing times than Kenyan runners
- ◯ None of the above

Q2.4 (2 points) Assuming the median (50th percentile) finishing time for Kenyan runners is 9,000 seconds, which of the following statements about Kenyan Runners is true?

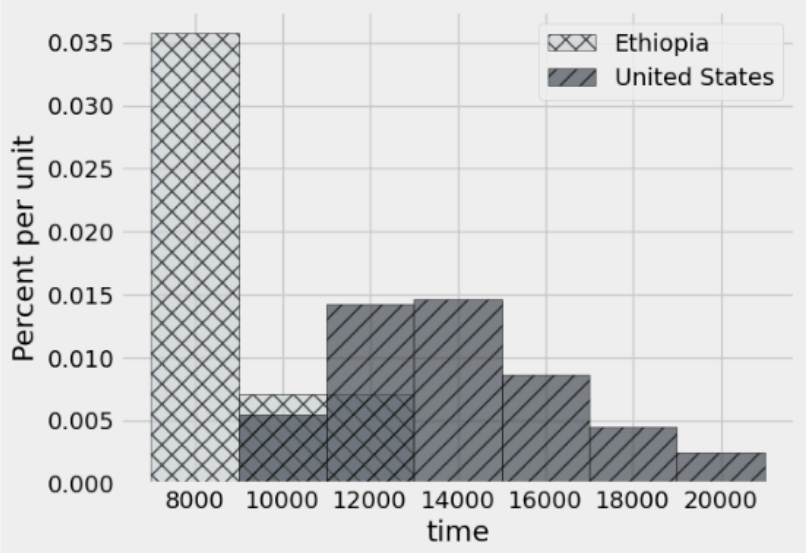
- ◯ At least 50% of runners finished in 9,000 seconds or less
- ◯ At most 50% of runners finished in 9,000 seconds or less
- ◯ 50% of runners finished in less than 9,000 seconds
- ◯ None of the above

Q2.5-2.7 (4 points) Based on the box plot, let's examine the spread of runners from the United States and Ethiopia more closely, since those box plots look quite different! Fill in the code below to produce the histogram shown. Assume **marathon_us_eth** has already been assigned to a table containing only runners from the United States and Ethiopia. Note that the bins range from 7,000 to 22,000 and are spaced evenly in increasing intervals of 2,000.

You do not need to replicate the color or shading shown-it is included only because the exam will be printed in black and white.

```
marathon_bins = _____
                  Q2.5

marathon_us_eth.hist(_____, bins = marathon_bins,
                       Q2.6
_____)
  Q2.7
```



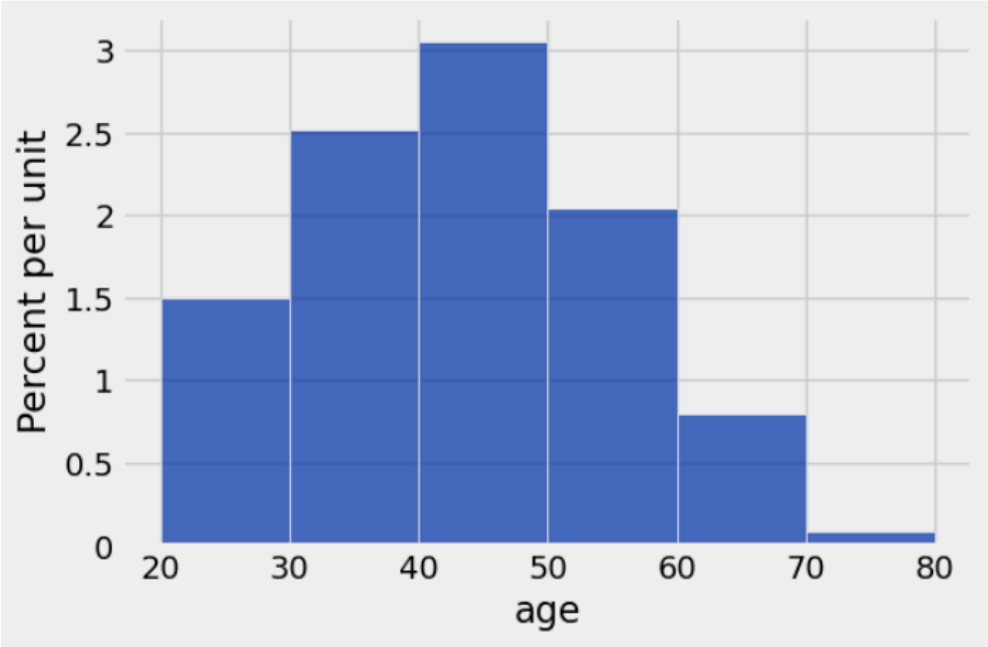
Q2.8 (1 point) Based on the histogram above, what is the "unit" in percent per unit?

- ○ Percent
- ○ Second
- ○ None of the above
- ○ Runner
- ○ Age

Q2.9 (2 points) Based on the histogram above, which of the following statements are true?

- ☐ Both histograms for the United States and Ethiopia are left-skewed
- ☐ Both histograms for the United States and Ethiopia are right-skewed
- ☐ This histogram is not a valid comparison of distributions because there are likely different numbers of runners in each group
- ☐ The average finishing time for United States runners is higher than the average finishing time for Ethiopian runners
- ☐ None of the above

Let's look at the data for all runners, regardless of country. The following histogram shows the distribution of age among all runners.



We will use the histogram to answer the next four questions.

Q2.10 (2 points) Based on the histogram above, which bin contains the **median** age?

- ◯ [30, 40)
- ◯ [40, 50)
- ◯ [50, 60)
- ◯ [30, 40]
- ◯ [40, 50]
- ◯ [50, 60]

Q2.11 (1 point) The mean age of runners is higher than the median age because the distribution is left skewed.

- ◯ True
- ◯ False

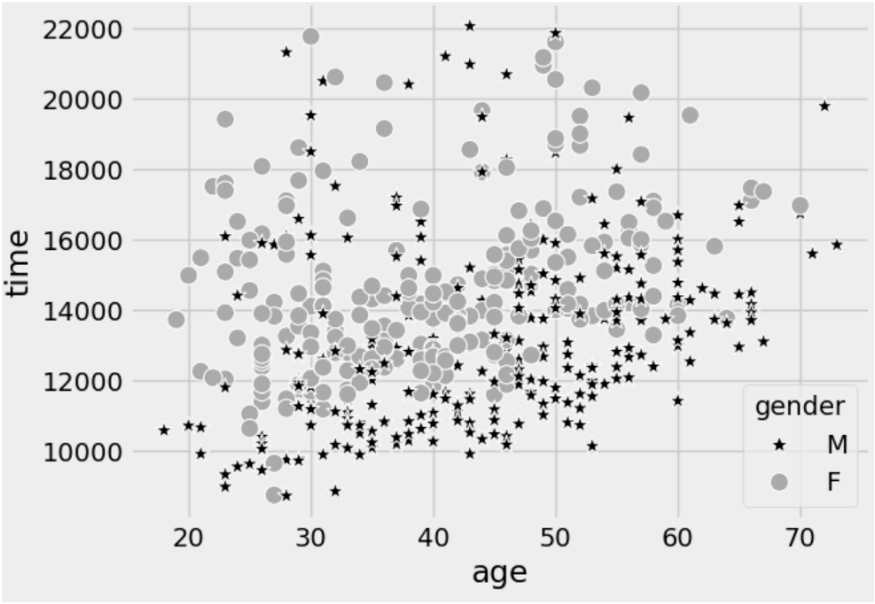
Q2.12 (1 point) More than 60% of runners in this dataset were between the ages of 30 and 50.

- ◯ True
- ◯ False

Q2.13 (2 points) What *percentage* of runners are younger than 40? Include your work in the box below and circle your final answer. Your final answer should be simplified. *Hint: The area of each bar can be calculated with* ***height * width****.*

(Question 2 continued...)

For the remaining questions, consider the scatter plot below. To avoid *overplotting* we've randomly sampled 500 runners from the overall dataset.



Q2.14 (1 point) How many variables are encoded in this visualization?

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4

Q2.15 (1 point) What is the **correlation** between age and finishing time?

- ◯ Positive
- ◯ Negative
- ◯ No correlation

Q3 *Just for fun!* **(0 points)**

Q3.1 Draw something fun, or write a message for the staff! Or leave this blank!